MATH 265

Spring 2019

Lab 4

Understanding the Page Rank Algorithm

Jacob Swift

Jacob Elenbaas

Nate Goodman

Sean Gordon

Patrick Perkins

May 28, 2019

Problem Motivation and Overview

Humanity has come a long way through their technological advances throughout the course of history, and we have found that comparatively few events that occur in nature can be controlled by us. Because of this lack of control, there will be probability: the possibility of varying outcomes, and a variety of them. In order to track this, we need to understand the concept of probability and how multiple events with differing probabilities interact with one another.

Simple probability calculations in the short term are not enough in certain contexts, and more advanced means are required in order to calculate the collective probabilities of long strings of events. These ideas apply to many occurrences in human life; from predicting the chances of certain types of weather over a long span of time to tracking the movement of a web surfer as he clicks from page to page. Nevertheless, this concept applies to countless situations and therefore is essential to understand if one is to gain an understanding of probability and how it can be estimated over a long amount of time.

Markov Chains are an example of a system used in internet page rank systems, and were developed around 1998. Hyperlinks concealed in web pages contributed to the ranking system, the display of the web page importance, adjusting the presumed path of the user. Search engines, like Google, depend on this page-rank algorithm to provide relevant web pages based on prior search history of each individual user. Taking into account time spent on each page, hyperlink topic relativity and a series of several other web page components helped Google create one of the world's most used search engines.

Introduction[1]

To begin this lab, we first need to know what we are being asked to do in each step. In 9.1 we want to see what happens to T as n approaches infinity and see if there is a limit. We check this by using Markov process. In a Markov process, each generation only depends on the generation directly before it to calculate the probability in a given system. Then for 9.2 were asked to prove that 1 is an eigenvalue for the eigenvector T. We can do this

[1] Intro/Conclusion adapted from http://www.ohiouniversityfaculty.com/mohlenka/goodproblems/intros.pdf

using the Equilibrium vector, which is Tx = x. Next, in 9.3 we want to use the steady-state vector and use it to find which page is the most important, now knowing that website 3 randomly redirects to any of the other page. With webpage 3 now able to continue surfers on to other pages, the hanging page problem is fixed. In 9.4 we look at a new transition matrix to see a different scenario of web pages, with a different size and different paths between the sites. We are also tasked with finding a steady-state vector for this internet. Once again we will be using the Markov process and this time we will add to our dangling pages rule from the start. Lastly, in 9.5 we looked at the matrix in 9.1 as a google matrix, finding a matrix and steady state vector for that version as well as finding the relative importance and time spent on each page.

Solution

(9.1) The equation converges on the third page, because there is no output for 3. Every page will eventually reach page three and become stuck on page three as n approaches infinity. To determine the various steps for the web surfer described in the lab, we began by taking the starting page, 6, and using its corresponding vector to find x_1 using $x_1 = T^1x_0$. By performing this process repeatedly, we can calculate the probability that the surfer will end up at a certain page at any number of clicked links after they start on page 6. After we began to understand what was being simulated in this experiment, we multiplied x_0 by the transformation T repeatedly. We began with T^1x_0, which gave us this vector:

$$[0 \quad 0 \quad 1/3 \quad 0 \quad 1/3 \quad 0 \quad 1/3]^T$$

After this, we continued multiplying by the transformation in order to determine the end behavior of the transformation after applying it many times. After we reached extremely high numbers of T applications, we noticed a pattern in the outcome.

$T^{10000}\, x_0 = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]^T$

We chose another starting page and calculated T^1x_0. The starting page we chose was 5. This calculation gave us this vector:

$$[0 \quad 0 \quad 0 \quad 1/2 \quad 0 \quad 1/2 \quad 0]^T$$

We then performed the same calculation for very large values of n. This gave us the following output:

$T^{10000}\, x_0 = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]^T$

We can see from these results that repeated use of the transformation is causing the outcomes to converge on page 3. After reapplying this outcome to the context of the experiment, we were able to determine exactly why this happens. It occurs because of the way that the web of pages is set up: 3 is the only page that contains no links that redirect to a different page, and therefore has no exit. When a web surfer reaches page 3 by random chance, it becomes impossible for them to leave page 3. Through the randomness of the surfing, we can conclude that almost every surfer will inevitably end up on page 3, which explains the convergence that can be observed with repeated uses of the transformation T.

Taking this pattern extremely far, while taking note of the convergence of the function, it can be observed that the vector output to the transformation always approaches $(0\ 0\ 1\ 0\ 0\ 0\ 0)^T$. This vector can otherwise be known as the *Equilibrium Vector*, and it is the destination upon which the repeated application of transformation T will lead vector x_n.

(9.2) Due to the nature of acquiring our Equilibrium vector, repeated multiplication of transformation T will not cause it to change, as this vector is effectively the point of convergence for $T^{n+1}x_n$, and therefore the input vector and output vector will be equal after we reach this point of equilibrium. This relationship that the function demonstrates after reaching this point can be summarized in terms of Eigenvalues and Eigenvectors.

$T^{n+1}x_n = x_{n+1}$

The above equation represents the process of the transformation T between x_n and the next vector in the sequence, x_{n+1}. At the point of Equilibrium, however, the function's output and input vectors will theoretically be equal. Because of this, we can rewrite the equation in a much simpler way.

$Tx = x$

Through this equation, the eigenvalues and vectors of the transformation can be calculated, but only when using the Equilibrium vector, as this vector is what the above equation is based around. When using this vector, because Tx=x, Tx=1x. Therefore, 1 is the Eigenvalue for T and x is its corresponding Eigenvector when using the Equilibrium vector. We can also show that the limiting vector is an eigenvector using the equation $(A - \lambda I_7)x = 0$ where $\lambda=1$. Observe:

$$\left(\begin{bmatrix} 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/2 & 0 & 1/3 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/3 & 1 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \end{bmatrix} - 1 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}\right) \mathbf{v} = 0$$

$$\begin{bmatrix} -1 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/3 & 0 \\ 0 & 1/2 & 0 & -1 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & -1 & 1/3 & 1 \\ 0 & 0 & 0 & 0 & 1/2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/3 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \\ v5 \\ v6 \\ v7 \end{bmatrix}$$

RREF Matrix

$$\left[\begin{array}{ccccccc|c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

This means that $v_1 = 0$, $v_2 = 0$, $v3 = v_3$, $v_4 = 0$, $v_5 = 0$, $v_6 = 0$, $v_7 = 0$. So we end up with this eigenvector:

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

So as it turns out, the resulting eigenvector is the equilibrium vector.

(9.3) In the previous parts of the experiment, we deduced that page 3 was the inevitable destination for all web surfers after they reached the vector of convergence, otherwise known as the Equilibrium vector. This was because page 3 had no exit, and therefore collected the surfers as they wandered without any possibility of exit. To circumvent this, we will add a 1/7 probability of moving to a page randomly from page 3. Our new matrix, which we will call A, will look like this:

$$\begin{bmatrix} 0 & 1/2 & 1/7 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1/7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/7 & 1/2 & 0 & 1/3 & 0 \\ 0 & 1/2 & 1/7 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/7 & 1/2 & 0 & 1/3 & 1 \\ 0 & 0 & 1/7 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/7 & 0 & 0 & 1/3 & 0 \end{bmatrix}$$

In order to determine whether or not our change worked, our next step will be to apply the transformation repeatedly to x_0 in order to determine the end behavior of the transformation A. After doing so, we see that the transformation is not convergent upon one page as the last one was. This knowledge proves that the adjustment applied to transformation T was indeed a success. Through repetitive application of transformation, A, we can find that the vector upon which Ax converges is as follows:

9.3 b.

$$\begin{bmatrix} 0.0732 \\ 0.0976 \\ 0.1707 \\ 0.1951 \\ 0.2439 \\ 0.1463 \\ 0.0732 \end{bmatrix}$$

This result proves that though we successfully removed the effective pitfall that was page 3, the probability of ending up on particular pages still varies due to the layout of the pages and the varying numbers of entrances and exits to each page.

9.3 c.

The most important page based on this new vector is page 5, but this is to be expected from such a varied system of pages such as this. Due to the layout of the pages, not all pages have equal exits and entrances, and as can be seen in the diagram, page 5 has 3 entrances and 2 exits, meaning that the ratio of entrances to exits is highest for page 5. This is significant because it means that the traffic in will be theoretically higher than the traffic out, which explains why the matrix A's convergence highlights 5 as the page containing the most web surfers.

(9.4) In this new situation we have a system with 5 pages, and a matrix that corresponds with this new system, matrix P. This matrix is a 5x5 matrix, which corresponds to the 5 separate pages. In each row, there are numbers that correspond to each page's possible exits, corresponding to the likelihood that a surfer will leave to a certain page from another specific page. The columns are composed similarly to matrix T and A from the previous problems. The columns represent page n, and the numbers within represent the traffic that is entering page n from page m, where m is the row in which the number is located. For example, in column 3 there is a "1" in row 1. This means that all 100% of traffic on page 3 will move to page 1, which we can clearly see is also true on the diagram. This comparison can be made in all columns of matrix P, another example being column 5, with the five ⅛ entries portraying the fact that when a surfer reaches 5 they will be sent to a random page with equal

likelihood. Because all columns in matrix P correspond to one of the pages in the diagram, it can be concluded that this matrix is indeed an accurate description of the diagram.

Now, similarly to the matrices we studied beforehand, we must examine matrix P's end behavior in order to obtain the point of convergence: the Equilibrium Vector. Using the same method as we used in the previous segments of the lab, we multiplied matrix P repeatedly by three separate iterations of x_0, as follows.

$$P^{1000}x_0 = [1/3 \quad 1/3 \quad 1/3 \quad 0 \quad 0 \quad 0 \quad 0]^T$$

We concluded in our experiment that the point of convergence does not depend on the starting vector x_o, but is instead based upon the configuration of pages and how many exits and entrances they have. For example, this transformation converges on an equal probability amongst pages 1-3, with a probability of 0 for pages 4 and 5, which suggests that the first three pages create some sort of self-contained space from which no surfer is able to leave, while they are able to enter from pages 4 and 5. This conjecture based on the Equilibrium vector is true, as the diagram shows, as pages 4 and 5 can send surfers to pages 1, 2 and 3 with no possibility of return, where surfers move cyclically through the aforementioned 3 pages in an infinite loop, confining them to only those 3 pages.

(9.5) We began this step by developing an understanding of what exactly we were being asked to do. We began by amassing the variables we knew. Since the problem referred to the matrix we were solving for as a "Google matrix" we knew that we had to use the Google example for p, which is stated in the book as being 0.85. Secondly, due to our usage of the 7-page internet from 9.1, we knew that n equals 7. With this information, we were able to solve the problem, as follows.

G = pT + (1 - p)Q

G = .85T + .15Q

$$G = .85\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 1 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & \frac{1}{7} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{7} & \frac{1}{2} & 0 & \frac{1}{3} & 1 \\ 0 & 0 & \frac{1}{7} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{3} & 0 \end{bmatrix} + .15\begin{bmatrix} \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{3}{140} & \frac{25}{56} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{25}{56} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{25}{56} & \frac{3}{140} & \frac{25}{56} & \frac{3}{140} \\ \frac{3}{140} & \frac{25}{56} & \frac{1}{7} & \frac{3}{140} & \frac{25}{56} & \frac{3}{140} & \frac{3}{140} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{25}{56} & \frac{3}{140} & \frac{25}{56} & \frac{25}{56} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{25}{56} & \frac{3}{140} & \frac{3}{140} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{25}{56} & \frac{3}{140} \end{bmatrix}$$

We used fractions to eliminate rounding errors and make calculations as accurate as possible. After calculating matrix G, we multiplied it by x_0 ad nauseam in order to determine its end behavior. After multiplying by G extensively, we calculated that it was indeed converging. We then found the equilibrium vector:

$$\begin{bmatrix} 0.0910 \\ 0.1181 \\ 0.1594 \\ 0.1875 \\ 0.2271 \\ 0.1373 \\ 0.0797 \end{bmatrix}$$

Finally, we can use this vector for real-world applications. This vector represents the relative rank for the pages, due to its intrinsic relation of all entries to the total, which is 1. Because the entries in this vector are all

probabilities, they are already relative to one another. The pages are listed here, in order of minimum to maximum importance in regards to rank:

Rank 1 ---------- Page 5: 0.2271

Rank 2 ---------- Page 4: 0.1875

Rank 3 ---------- Page 3: 0.1594

Rank 4 ---------- Page 6: 0.1373

Rank 5 ---------- Page 2: 0.1181

Rank 6 ---------- Page 1: 0.0910

Rank 7 ---------- Page 7: 0.0797

When working in terms of time, we can make identical conjectures about this set of numbers, as the numbers also represent time. When examining the probabilities above, it is given that they represent the probability of a web user being on said page at any given time. However, the probability of a user being on a page also dictates the percentage of time spent on said page. The reason for this is due to the linear relationship between occupation and time - if a space has a 22% chance to be occupied at any given time, 22% of the time will be spent with the space being occupied. The relationship between the two always works in this way, and therefore can be interpreted either way. Knowing this fact, we can conclude that the amount of time spent on a given page will be equivalent to the probability a user ends up on the aforementioned page.

Conclusion

Once we got the hang of the various concept that this lab introduced, it was generally easy to follow along with each step of the process. In part 9.1, we got an understanding of what the concepts in the lab were. For starters, we figured out what x_0 represented, how the matrix worked with it, and what the different parts meant in the context of a long term and constantly changing probability based on numerous external factors. After we did this, we solved the problem by effectively finding the end behavior, the Equilibrium Vector as it is referred to in

the lab, by multiplying the transformation matrix by the starting vector repeatedly. We found that the result converged upon x_0, which made sense due to the fact that page 3 had no exit. It was only reasonable that the web surfers would eventually end up trapped there, as the probability of landing on page 3 increased with every multiplication of the matrix. In 9.2 we considered this situation in terms of eigenvalues and eigenvectors. We concluded, due to the fact that the equilibrium vector multiplied by the matrix equaled the equilibrium vector, that the eigenvalue of the transformation vector is 1, and its corresponding eigenvector is the equilibrium vector. In 9.3, we applied the requested change to matrix T, and made sense of the changes this caused in the behavior of the vector. The change we made caused the most important page to shift away from 3, as it was no longer the point of convergence due to the newly created exits from said page. By multiplying the starting vector by the matrix repeatedly, we found that the equilibrium vector highlighted 5 as the most important page, but this was easily explainable due to the fact that 5 had a higher ratio of entrances to the page when compared to exits. In 9.4, we first explained the connection between the given 5x5 matrix, which we named matrix P. After this, we justified the fact that the starting vector did not affect the equilibrium vector for the matrix. We proved this through the fact that end behavior is depended on the matrix and not the function, and no matter where the vector begins, the matrix will emulate the same pattern and exhibit the same type of end behavior. In 9.5, we used external variables to convert our matrix in 9.1 to Google matrix G. We then found its equilibrium vector and used it to create a list of the different pages, their respective probabilities and their ranks, in order. Following this, we justified the idea that page-rank and time were both defined by the same probability, because the chance of a user being on a page and the percentage of time spent on that page are very interwoven in concept.

Reflection of Solution Process

To conclude this lab, we learned a great deal about the inner workings of probabilities and the way they can be portrayed through vectors and matrices, as well as what happens when these probabilities are multiplied by one another, and how to calculate the outcomes for extremely long term chains of different probabilities. This

knowledge is valuable for application in a countless number of fields, with a prime example being the prediction of weather in weather forecasts. An example of a small challenge we faced and overcame came was with the difficulty we had with comprehending how exactly the large 7x7 matrix was related to x_0 in the first part of the lab. We began by looking carefully at the matrix itself, thinking that x_0 was somehow derived from one of its columns or rows. After being confused by this, we went back and read the instructions more carefully. We realized that x_0 had no relation to the entries of the matrix, but was instead the starting point for the series of transformations done by multiplying by the matrix. This moment helped us collectively begin to understand exactly how the probabilities worked, and that x_0 was simply the start to the series that was created by multiplying it by the transformation matrix repeatedly. One of our most useful tools in the development of the lab was without a doubt GeoGebra. This site helped us immensely with determining the effects of the matrix on different vectors in a more efficient way than calculators. This was mostly because calculators take a lot of effort to type out larger matrices and test them with different vectors, while this website allowed us to easily change the vectors at hand, and study how the matrices' effects on them differ. The aforementioned calculators were very ineffective, however, as we had tried to use them mostly before we started using GeoGebra. Our calculators proved to simply be too slow to suit our needs, in addition to this GeoGebra was easier to show off to the entire group, so it provided an easy method to keep up on our progress during the lab.

References

Austin, D. (n.d.). How Google Finds Your Needle in the Web's Haystack. Retrieved from http://www.ams.org/publicoutreach/feature-column/fcarc-pagerank

"Chapter 10: Finite-State Markov Chains." *Www.faculty.winthrop.edu*, faculty.winthrop.edu/polaskit/spring11/math550/chapter.pdf.